Peak-integration uncertainties for HR–AMS—PMF: overlapping peaks & importance of mass accuracy

When do we need uncertainties?

- PMF
- ME-2
- Linear regression (e.g. $C_2H_3O^+$ vs. $CO_2^+$)

Both PMF & linear regression solve

$$y = FG + e$$

\[\text{data} = \text{model} + \text{residuals}\]

by minimizing

$$Q = \sum_{i,j} \left( \frac{e_{ij}}{\sigma_{ij}} \right)^2$$

...which makes the precision (imprecision $\sigma$) as important as the measurements ($y$).\(^1\)

Current HR-AMS-peak uncertainties

Same as for UMR-AMS:

“how would this signal vary if we repeated the measurement?”

\[ \sigma_n = \sqrt{n_{\text{ions}}} \]

Needs “corrections” for sampling configuration (sampling time, etc.)
see Allan et al., JGR 2003
or the AMTD link on Slide 1
Another source of uncertainty:

“How would this peak-integration vary if we repeated the analysis?”

→ Rest of this talk
Outline

1. PIKA peak-integration errors

2. Understanding of peak-fitting errors via case study for single (isolated) ions

3. Application to single & overlapping peaks

4. Practical application

NB: For simplicity I won’t discuss the PIKA uncertainty changes in 1.11C. They make no difference in this context.
Definitions

• \( \sigma \) = imprecision

• \( \varepsilon \) = imprecision + bias (overall error)

• * All existing peaks assumed to be identified
Essential ideas in PIKA

1. Fit height $h$ of pseudo-Gaussian:

$$f(x) = h \cdot v \cdot \exp\left(\frac{(x - \mu)^2}{-w^2}\right)$$

$$f(x) = hf_0$$

2. Integrate fitted peak:

$$A = hw\sqrt{\pi}\left(k_{DC} \frac{A_{\text{pseudoG}}}{A_{\text{Gauss}}}\right)$$

3. Estimate error:

$$\sigma_n = \sqrt{n}$$

$$\left(\frac{\sigma_A}{A}\right)^2 = \left(\frac{\sigma_h}{h}\right)^2 + \left(\frac{\sigma_w}{w}\right)^2$$

**Peak-integration uncertainty**

*counting imprecision*
Estimating $\sigma_A$

\[
\left(\frac{\sigma_A}{A}\right)^2 = \left(\frac{\sigma_h}{h}\right)^2 + \left(\frac{\sigma_w}{w}\right)^2
\]

\[A = hw\sqrt{\pi} \left( k_{DC} \frac{A_{\text{pseudoG}}}{A_{\text{Gauss}}} \right)\]

$\sigma_{\text{AMS}} = \sigma_n + \sigma_A$

$\sigma_w \leftarrow$ direct estimate from width calibration

$\sigma_h \leftarrow$ Monte-Carlo estimate

from empirically-estimated $\sigma_v$, $\sigma_w$, $\varepsilon_\mu$
Outline

1. PIKA peak-integration errors

2. Understanding of peak-fitting errors via study of 7 single (isolated) ions in test data set

\[ \sigma_h \leftarrow \text{Monte-Carlo estimate} \]

from empirically-estimated \( \sigma_v, \sigma_w, \varepsilon_\mu \)
Fit errors (RMSE’s) for 7 isolated peaks

Each point shows a peak.

All ions show:
1. Noise regime < 10 cps
2. Constant %RMSE regime

Constant ratio because $h$ scales errors in $f_0$:

$$f(x) = hf_0$$

(details in AMT paper)

These are Diff data to avoid differences in backgrounds.
What controls the %RMSE?

- Only noise in location $\mu$ matters for fitting errors.
- If %RMSE increases, red data become higher than blue data.
- Noise added within PIKA (amount of noise chosen to have visible effects).
- Errors ~10x higher.
- Errors ~100x higher.
- Only noise in location $\mu$ matters for fitting errors.
\(m/z\) prediction errors \(\varepsilon_\mu\) are the major cause of fitting errors (RMSE).

Location-prediction \((m/z\) calibration\) errors \(\varepsilon_\mu\) are \(\sim10\times\) more important than peak-shape errors.

\[
f(x) = h \cdot v \cdot \exp\left(\frac{(x - \mu)^2}{-w^2}\right)
\]

\[
f(x) = hf_0
\]
Estimating $\sigma_A$ (updated)

\[
\left(\frac{\sigma_A}{A}\right)^2 = \left(\frac{\sigma_h}{h}\right)^2 + \left(\frac{\sigma_w}{w}\right)^2
\]

\[\sigma_{\text{AMS}} = \sigma_n + \sigma_A\]

$\sigma_w \leftarrow$ direct estimate from width calibration

$\sigma_h \leftarrow$ Monte-Carlo estimate from empirically-estimated $\sigma_p, \sigma_w, \varepsilon_{\mu}$
So, we need to know $\varepsilon_\mu$: for isolated peaks, estimate as

$$\text{Mean difference: } (\mu_{\text{free-fit}}) - (\mu_{m/z \text{ cal}})$$

Different bias (mean) and imprecision (spread) at each ion, even for $\text{C}_2\text{H}_3\text{O}^+$ and $\text{C}_3\text{H}_7^+$ ($m/z 43, \Delta m/z = 0.036$).

Can only estimate ion-specific $\varepsilon_\mu$ for isolated peaks!
Outline

1. PIKA peak-integration errors

2. Understanding of peak-fitting errors*
   – via case study for single (isolated) ions

3. Application to single & overlapping peaks:
   – Monte Carlo estimation

4. Practical application
Monte Carlo $\sigma_h$ estimation

1. Estimate $\mu_i, w_i$
2. Add estimated error to $\mu_i$ ($\mu_{fit} = \mu_i + \varepsilon_\mu$)
3. Do fit; save $h_{fitted}$
4. Repeat $n$ times

$h = \text{Avg}(h_{fitted})$
$\sigma_h = \text{SD}(h_{fitted})$
Monte Carlo $\sigma_h$ estimation

1000 fits to $C_3H_7^+$

Errors in $\mu$ programmed as gaussian mean and variance.
On zoomed graphs, middle line is bias, other lines are bias $\pm 1\sigma$ imprecision.
## Summary of all 7 ions’ $\sigma_h$

<table>
<thead>
<tr>
<th>Ion</th>
<th>$\mu$-prediction error</th>
<th>Error in fitted $h$ (bias [%], imprecision $\sigma_h$ [%]) from best-estimate $\mu$ errors with $\mu$ imprecision only with broader peaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>C$_2$H$_3^+$</td>
<td>4.6 ± 9.5</td>
<td>−0.35, 1.06</td>
</tr>
<tr>
<td>C$_3$H$_7^+$</td>
<td>−10 ± 7.5</td>
<td>−0.65, 1.64</td>
</tr>
<tr>
<td>C$_4$H$_2^+$</td>
<td>−14 ± 5.7</td>
<td>−1.06, 2.10</td>
</tr>
<tr>
<td>C$<em>5$H$</em>{11}^+$</td>
<td>−15 ± 5.0</td>
<td>−1.03, 2.46</td>
</tr>
<tr>
<td>C$_3^+$</td>
<td>4.8 ± 12</td>
<td>−0.51, 1.52</td>
</tr>
<tr>
<td>C$_2$H$_3$O$^+$</td>
<td>−5.9 ± 8.3</td>
<td>−0.39, 1.36</td>
</tr>
<tr>
<td>CO$_2^+$</td>
<td>6.9 ± 9.0</td>
<td>−0.12, 1.41</td>
</tr>
</tbody>
</table>
How to deal with “unknowable” $\varepsilon_\mu$?

$\varepsilon_\mu = \text{bias}(\mu) + \sigma_\mu$

at A: $\sigma_\mu = 5\%$, neglecting a bias($\mu$) of ±10 strongly affects $\sigma_h$ (1–2%) - Overestimate $\sigma_\mu$ !!

at B: $\sigma_\mu = 15\%$, neglecting a bias($\mu$) of ±10 negligibly affects $\sigma_h$ (3%)
Cross-section of previous image (3rd row)

\[ \varepsilon_\mu = \text{bias}(\mu) + \sigma_\mu \]

\( \sigma_\mu = 5\% \), neglecting a \( \text{bias}(\mu) \) of ±10 strongly affects \( \sigma_h \) (1--2\%)

\( \sigma_\mu = 15\% \), neglecting a \( \text{bias}(\mu) \) of ±10 negligibly affects \( \sigma_h \) (3\%)

Overestimate \( \sigma_\mu \)!!
Application to overlapping peaks:

MC with 100 samples, exactly the same scenario.

*Inputs:* a single $\sigma_{\mu}$ estimate + the usual peak parameters ($w, \nu, \mu$)
This matters for PMF

- Test data set with high signals
  - Counting imprecision
  - 5% imprecision (not overlapping errors)

- Without including 5% in PMF, high signals became residual spikes

- Lower-signal factors retrieved with $r^2 0.54$ without 5% included in $\sigma$, down from 0.74
  - Higher signals from 0.91 to 0.99

- Details in AMTD
Practicalities in $\sigma_h$ estimation

- MC estimation with 100 fits takes 100x longer.

- For preliminary analyses, we could therefore:
  - Estimate the uncertainty in isolated peaks for only a subset of peaks, since $\sigma_h/h$ is roughly constant [1]
  - Use the Cubison and Jimenez parameterization [2] to estimate the uncertainty in peaks that overlap significantly
    - For multiple overlapping peaks, we consider only the two closest peaks

- At the final stage of analysis, these initial estimates can be refined by MC estimation. The refined estimates will have the advantages of accounting for possible m/z-axis sensitivities, and for cases where >2 peaks overlap significantly.

- (Until this is available in PIKA, PMF’s C3 parameter provides a rough approach to isolated-peak errors)

Summary

1. Peak-integration uncertainties dominate ions with high signals or bad overlap.
   – For well-resolved, high-signal ions, the uncertainties can be estimated as constant %.

2. These can be quantified by Monte Carlo with $m/z$ cal. imprecision $\sigma_\mu$ as the only input.
   – Applies to overlapping & isolated ions (gives the constant% above).
   – The range of reasonable $\text{bias}(\mu)$ needs to be roughly known, so that $\sigma_\mu$ can be increased to account for the effect of $\text{bias}(\mu)$ on $\sigma_h$.

3. Will be incorporated into PIKA soon.